

# **FINDMODEL: A Tool to Select the Best-Fit Model of Nucleotide Substitution**

by

**Ning Tao**

B.E., Civil Engineering, Beijing Polytechnic University, 1998

THESIS

Submitted in Partial Fulfillment of the  
Requirements for the Degree of

Master of Science  
Computer Science

The University of New Mexico

Albuquerque, New Mexico

July, 2005

©2005, Ning Tao

# Acknowledgments

I would like to express my deepest gratitude to my thesis supervisors, Dr. Bernard Moret, Professor of the Department of Computer Science at University of New Mexico and Dr. Carla Kuiken, Technical Staff Member of Theoretical Biology Biophysics Group (T-10) at Los Alamos National Laboratory, for their continuous guidance, encouragement, and support through the course of this study.

# **FINDMODEL: A Tool to Select the Best-Fit Model of Nucleotide Substitution**

by

**Ning Tao**

## **ABSTRACT OF THESIS**

Submitted in Partial Fulfillment of the  
Requirements for the Degree of

Master of Science  
Computer Science

The University of New Mexico

Albuquerque, New Mexico

July, 2005

# FINDMODEL: A Tool to Select the Best-Fit Model of Nucleotide Substitution

by

**Ning Tao**

B.E., Civil Engineering, Beijing Polytechnic University, 1998

M.S., Computer Science, University of New Mexico, 2005

## Abstract

**Motivation:** Choosing a model of sequence evolution is a crucial step when using DNA sequence data to reconstruct phylogenies: using a mismatched model will reduce accuracy and may lead to erroneous conclusions. FINDMODEL is a web-based tool for selecting a model of DNA (nucleotide) evolution; it is designed to be easy to use by researchers who do some sequencing and may not have access to phylogenetic packages.

**Approach:** FINDMODEL can analyze 28 models or a restricted subset of 12 models. It creates a guide tree using Weighbor, optimizes branch lengths, calculates the likelihood for every chosen model (using `baseml` from the PAML package), and computes the Akaike information criterion (AIC). The model with the smallest AIC score is considered to be the best-fit model. Because of server limitations, the FINDMODEL web server processes inputs above a certain size in non-interactive mode,

sending an email to the user when it has completed the analysis with user’s data and providing a down-loadable file with the results.

**Results:** To test the performance of FINDMODEL, we generated simulated DNA sequences with **Seq-Gen** under four different models of nucleotide substitution of different complexity and compared the inferred model with the true model. We used 17 different configurations, with 5 instances for each set of parameter values. FINDMODEL returned the correct model for 73% of our test instances, and for another 9% returned the correct model, but with variable site-specific rates instead of homogeneous rates. Moreover, on all tests where FINDMODEL did not return the correct model, the normalized AIC error between the correct and the predicted models was below 0.002 (and the actual AIC difference was below 7).

# Contents

List of Figures	x
List of Tables	xi
<b>1 Introduction</b>	<b>1</b>
1.1 Overview . . . . .	1
1.2 Models of DNA evolution . . . . .	3
1.2.1 Jukes-Cantor's one-parameter model . . . . .	3
1.2.2 Kimura's two-parameter model . . . . .	4
1.2.3 HKY model . . . . .	5
1.2.4 The general time-reversible model . . . . .	5
<b>2 Approach</b>	<b>7</b>
2.1 Process . . . . .	7
2.2 Tools . . . . .	8
2.2.1 fmtseq . . . . .	8

2.2.2	gapstrip . . . . .	8
2.2.3	PHYML . . . . .	9
2.2.4	Neighbor . . . . .	9
2.2.5	PAML (and baseml) . . . . .	9
2.2.6	AIC . . . . .	10
2.3	Some other features of FINDMODEL . . . . .	10
<b>3</b>	<b>Experimental Setup</b>	<b>15</b>
<b>4</b>	<b>Results</b>	<b>18</b>
4.1	Results Regarding Models . . . . .	18
4.2	Results Regarding Lengths of Sequences . . . . .	23
4.3	Results Regarding $\Gamma$ Rate Heterogeneity . . . . .	28
<b>5</b>	<b>Discussion</b>	<b>30</b>
	<b>References</b>	<b>32</b>



# List of Figures

1.1	One-parameter model of nucleotide substitution. In this model, the rate of substitution in each direction is $\alpha$ . From [11]. . . . .	4
1.2	Two-parameter model of nucleotide substitution. In this model, the rate of transition ( $\alpha$ ) may not be equal to the rate of each of the two types of transversion ( $\beta$ ). From [11]. . . . .	5
2.1	The home page of FINDMODEL web site . . . . .	13
2.2	One of the result pages of FINDMODEL web site . . . . .	14

# List of Tables

1.1	The general time-reversible model of DNA evolution [10] . . . . .	6
2.1	Models considered by FINDMODEL. Models in the reduced set are in bold and with reference. . . . .	11
3.1	Parameters for tests of the JC and K2P models . . . . .	16
3.2	Parameters for tests of the HKY model . . . . .	17
3.3	Parameters for tests of the GTR model . . . . .	17
4.1	Test results. Empty entries indicate perfect matches; an asterisk (*) indicates a match, but with a $\Gamma$ rate parameter added; a question mark (?) denotes a mismatch of low significance, where the AIC error rate ( <i>AIC difference between correct model and selected model / AIC of selected model</i> ) was below 0.0005 (all entries marked with an asterisk also met this criterion); finally, more significant mismatches are indicated by the erroneous choice of model and, in parentheses, the corresponding AIC error rate. (The AIC values obtained for each test are shown in Table 4.2 to Table 4.18) . . . . .	19

4.2	AIC values of JC model test 1. In this and all the following AIC value tables, the AIC values of selected models are in bold, and the ones of correct models are in italic. . . . .	20
4.3	AIC values of JC model test 2 . . . . .	20
4.4	AIC values of K2P model test 1 . . . . .	21
4.5	AIC values of K2P model test 2 . . . . .	21
4.6	AIC values of K2P model test 3 . . . . .	22
4.7	AIC values of HKY model test 1 . . . . .	22
4.8	AIC values of HKY model test 2 . . . . .	23
4.9	AIC values of HKY model test 3 . . . . .	23
4.10	AIC values of HKY model test 4 . . . . .	24
4.11	AIC values of HKY+ $\Gamma$ model test 5 . . . . .	24
4.12	AIC values of HKY+ $\Gamma$ model test 6 . . . . .	25
4.13	AIC values of HKY+ $\Gamma$ model test 7 . . . . .	25
4.14	AIC values of GTR model test 1 . . . . .	26
4.15	AIC values of GTR model test 2 . . . . .	26
4.16	AIC values of GTR+ $\Gamma$ model test 3 . . . . .	27
4.17	AIC values of GTR+ $\Gamma$ model test 4 . . . . .	27
4.18	AIC values of GTR+ $\Gamma$ model test 5 . . . . .	28

# Chapter 1

## Introduction

### 1.1 Overview

Phylogenetics is the study of the reconstruction of the evolutionary history of genes and organisms by a combination of molecular biology and statistical techniques [24]. It has become nearly ubiquitous in biological and biomedical research as well as an important area of research in computer science. Phylogenetic analysis of DNA sequences is a fundamental tool in the study of the evolutionary history of organisms, from bacteria to humans [14, 15, 26, 28, 40]. Molecular data, especially DNA sequence data, are much more powerful for evolutionary studies than data from some traditional means of evolutionary inquiry such as morphology and physiology for several reasons. First, DNA sequences often evolve in a more regular manner. Second, molecular data are more amenable to quantitative treatments and therefore can be used with sophisticated mathematical and statistical methods. Third, molecular data are much more abundant [10, 11, 24, 29].

The task of molecular phylogenetics is to convert information in sequences into an evolutionary tree for those sequences [29]. A great number of tree construction

methods have been proposed, since no single method performs well in all situations. The most popular methods can be classified into three types: distance matrix methods, maximum parsimony methods, and maximum likelihood methods. In distance matrix methods, the evolutionary distances, which are the numbers of nucleotide substitutions between all members of a set of sequences, are computed, then a phylogenetic tree is constructed. Maximum parsimony methods reconstruct the evolution of a site on a tree that requires the fewest evolutionary changes. Maximum likelihood methods choose the tree (or trees) that of all trees is the one that is most likely to have produced the observed data.

Models of sequence evolution, which make assumptions about the process of nucleotide substitution, play an important role when using DNA sequences to estimate phylogenetic relationships among organisms [20, 25, 31, 32]. Models will be explained in more detail in the next section. Different data sets are often best explained by different models—no single model fits every data set. The use of inappropriate models of phylogenetic analysis may result in less accurate or even erroneous conclusions, since the estimates of branch lengths and topology can be severely affected [4, 7, 17, 23, 37, 38]. Model selection is not only important in phylogenetic analysis, but also for estimating substitution parameters or for hypothesis testing [1, 31, 41, 43, 46, 47]. Yet models are often used blindly in analysis [31]: a specific model is often used either because it has been used by other authors or because it is the default option in the analysis package.

The best-fit model for a particular data set can be selected through statistical testing. *Model selection* aims to find the model that most accurately estimates the unknown model of molecular evolution, while avoiding bias and excessive variance [25]. Study results suggest that model selection is reasonably accurate [32].

The software ModelTest [30], written specifically for testing whether the chosen model is appropriate, can also be used for model selection. However, ModelTest

requires access to PAUP\* [39], which, while a standard package for phylogenetic analysis, does not run under Windows and is aimed at expert users. In response to the need for a user-friendly tool aimed specifically at model selection, we developed FINDMODEL. FINDMODEL is web-based and thus accessible from any platform; it includes 28 different models of nucleotide evolution.

The functionality of FINDMODEL, and the methods and phylogenetic packages used by it are described in chapter 2. In chapter 3, we explain the experimental setup for testing its performance. Chapter 4 discusses the experimental results. Chapter 5 discusses the alternative methods for model selection and future work.

## **1.2 Models of DNA evolution**

The change in nucleotides with time is essential for understanding the evolution of DNA sequences and is used both in estimating the rate of evolution and in reconstructing the evolutionary history of organisms [10, 11, 24, 29]. Many models have been proposed for studying this process. We explain here in some detail the models we used to test the performance of FINDMODEL. More details about these models and about many other models can be found in the referred books and articles.

### **1.2.1 Jukes-Cantor's one-parameter model**

The model of Jukes and Cantor is the simplest model of DNA sequence evolution [16]. The substitution scheme it uses is shown in Figure 1.1. This model assumes equal chance of changing for each base in the sequence and no bias in the direction of change. This results in an equal frequency of the four bases at equilibrium. In this model, the rate of substitution for each nucleotide is  $3\alpha$  per unit time, and the rate of substitution in each of the three possible directions of change is  $\alpha$ . This model is

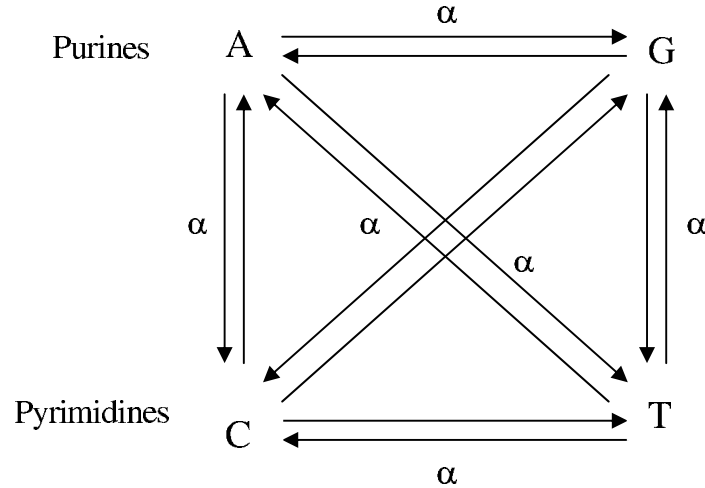


Figure 1.1: One-parameter model of nucleotide substitution. In this model, the rate of substitution in each direction is  $\alpha$ . From [11].

also called the one-parameter-model, since it involves only one parameter.

### 1.2.2 Kimura's two-parameter model

In most cases, nucleotide substitutions do not always occur randomly, as assumed in the Jukes and Cantor model. Kimura introduced a two-parameter model that allows a transition/transversion inequality of rate [18]. The substitution scheme is shown in Figure 1.2. In this scheme, the rate of transitional substitution at each nucleotide site is  $\alpha$  per unit time, whereas the rate of each of the two types of transversional substitution is  $\beta$  per unit time. The ratio of transitions to transversions will be  $\alpha/(2\beta)$ . The total rate of change will be  $\alpha+2\beta$ . This model is symmetrical, so the equilibrium frequencies of all four bases under it are also equal.

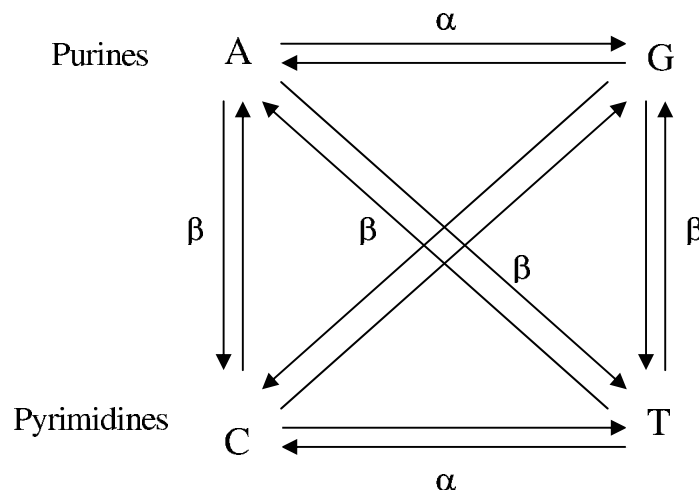


Figure 1.2: Two-parameter model of nucleotide substitution. In this model, the rate of transition ( $\alpha$ ) may not be equal to the rate of each of the two types of transversion ( $\beta$ ). From [11].

### 1.2.3 HKY model

The Kimura two-parameter model and the Jukes-Cantor one-parameter model both assume that all four bases have equal expected frequencies. The HKY model, which relaxes this assumption, was introduced by Hasegawa, Kishino, and Yano [13]. It extends the Kimura two-parameter model to asymmetric base frequencies and has five parameters.

### 1.2.4 The general time-reversible model

All the models mentioned above are reversible. When the equilibrium frequencies of the bases are  $\pi_A$ ,  $\pi_C$ ,  $\pi_G$ , and  $\pi_T$ , a model is reversible if

$$\pi_i \text{Prob}(j \mid i, t) = \pi_j \text{Prob}(i \mid j, t) \quad (1.1)$$



Table 1.1: The general time-reversible model of DNA evolution [10]

To: From:	A	G	C	T
A	-	$\pi_G\alpha$	$\pi_C\beta$	$\pi_T\gamma$
G	$\pi_A\alpha$	-	$\pi_C\delta$	$\pi_T\epsilon$
C	$\pi_A\beta$	$\pi_G\delta$	-	$\pi_T\eta$
T	$\pi_A\gamma$	$\pi_G\epsilon$	$\pi_C\eta$	-

In this case, if base  $i$  is at one end of a branch, and base  $j$  is at the other end, there is no way to decide which end was the ancestor and which the descendant, because the probability of starting with  $i$  at one end, and ending with  $j$  at the other, is the same as the probability of starting with  $j$  and evolving to  $i$ . Reversibility is the basic reason why we usually are not able to place the root of a tree. The instantaneous rates of change for this model for DNA are shown in Table 1.1 [22]. The  $\pi_i$  are the equilibrium frequencies of the bases, so the total rate of change will be the sum of the off-diagonal elements of the table, each multiplied by the probability that one would start with that base.

# Chapter 2

## Approach

FINDMODEL analyzes the input alignment to decide which of a predefined collection of models of character evolution best describes the input data, using an idea first implemented in ModelTest [30]. It is one of the applications of the Los Alamos hepatitis C sequence database (HCV database) [21] and is available at [hcv.lanl.gov/content/hcv-db/find\\_model/findmodel.html](http://hcv.lanl.gov/content/hcv-db/find_model/findmodel.html). Its homepage and result page are shown in Figure 2.1 and Figure 2.2.

### 2.1 Process

The input sequences are converted into FASTA format using `fmtseq` (see Tools, below). If the format of the input file is not recognized by `fmtseq`, FINDMODEL suggests other options for converting the sequences into FASTA format. Input sequences are then checked to ensure that they are legal nucleotide sequences. If the input file is above a certain size, FINDMODEL saves the input file, asks the user for an email address and for a title for the job, and proceeds through the steps listed below under control of a script, at the end of which it will email the results to the user.

All columns containing gaps are removed from the input alignments. **DNAdist** from the **PHYLIP** [9] package is used to create a distance matrix. **Weighbor** [5] is then used to reconstruct a tree from that distance matrix. At that point, each of the models in the chosen set (currently, a full set of 28 or a reduced set of 12) is evaluated in turn. To evaluate a model, **baseml** from the **PAML** package is used to optimize the branch lengths of the tree (the most expensive part of the computation), then the *Akaike information criterion (AIC)* [2] is calculated. The model with smallest AIC score is considered to be the best-fit model [30]. The program provides the user with the likelihood and AIC score for each model considered, plus the model selected and the values of its parameters.

## 2.2 Tools

### 2.2.1 fmtseq

**fmtseq** (available at [bioweb.pasteur.fr/docs/seqio/fmtseq\\_doc.html](http://bioweb.pasteur.fr/docs/seqio/fmtseq_doc.html)) is a reimplementation and extension of Gilbert's **readseq** program whose main function is to convert biological sequence files from one format to another. It recognizes formats Plain, EMBL, Swiss-Prot (**sprot**), GenBank (**gb**), PIR (**codata**), ASN.1 (**asn**), FASTA (Pearson), FASTA-old, FASTA-output (**fout**), BLAST-output (**bout**), NBRF, NBRF-old, IG/Stanford (**ig**), IG-old, GCG, MSF (**gcg-msf**), **PHYLIP**, **PHYLIP-Int** (**phylipi**), **PHYLIP-Seq** (**phylips**), **Clustalw** (**clustal**), and **Pretty**.

### 2.2.2 gapstrip

**gapstrip** (available at [hcv.lanl.gov/content/hcv-db/GAPSTRIP/gapstrip.html](http://hcv.lanl.gov/content/hcv-db/GAPSTRIP/gapstrip.html)), a locally developed script, removes any column in the alignment that contains one

or more gap characters and thus also reduces all sequences to the same length—that of the shortest sequence.

### 2.2.3 PHYLIP

PHYLIP (available at [evolution.genetics.washington.edu/phylip.html](http://evolution.genetics.washington.edu/phylip.html)) is a widely distributed phylogeny package written by J. Felsenstein. `dnadist` is one of the many programs available in PHYLIP; it uses nucleotide sequences to compute a distance matrix under one of four different models of nucleotide substitution. The default model is F84 [19] and is used to run `dnadist` in FINDMODEL. The pairwise distance for each pair of sequences is a maximum likelihood estimate of the divergence time (total branch length) between the two sequences. We chose to use `dnadist` because it is widely available, very well tested, and works well with Weighbor.

### 2.2.4 Weighbor

Weighbor (available at [www.t10.lanl.gov/billb/weighbor](http://www.t10.lanl.gov/billb/weighbor)) is a distance-based phylogeny reconstruction method. In effect, it is a weighted version of neighbor-joining [36] that gives significantly less weight to the longer distances in the distance matrix. The weights are based on variances and covariances expected in a simple Jukes-Cantor model. Weighbor is used in FINDMODEL because it is much faster than maximum likelihood, usually faster than maximum parsimony, and less sensitive than neighbor-joining to the presence of distant taxa.

### 2.2.5 PAML (and baseml)

PAML (available at [abacus.gene.ucl.ac.uk/software/paml.html](http://abacus.gene.ucl.ac.uk/software/paml.html)) is a package for phylogenetic analysis of DNA or protein sequences by maximum likelihood. `baseml`

carries out a maximum-likelihood analysis of nucleotide sequence evolution. The process of substitution is assumed to be stationary and Markov process models are used to describe substitutions between nucleotides. A discrete gamma model [45] is used to accommodate rate variation among sites. `baseml` can estimate tree topology, branch lengths, and substitution parameters, with a multitude of options, but it does not support invariant sites, in part because the estimate of the fraction of invariant sites tends to be very sensitive to the number of taxa. Since FINDMODEL uses `baseml`, it does not support invariant sites either—whereas ModelTest does, because PAUP\*, its phylogenetic reconstruction tool, can include an estimate of the number of invariant sites.

### 2.2.6 AIC

The Akaike information criterion is a measure of fit where the best fitting model is the one with the smallest AIC value. It is defined as

$$AIC = -2 \ln L + 2N \tag{2.1}$$

where  $L$  is the maximum likelihood for a specific model using  $N$  independently adjusted parameters within the model [2, 30]. AIC rewards models for good fit, but imposes a penalty for extra parameters, so fitting an excessively complex model is not likely [3, 12]. AIC allows for model selection uncertainty and model averaging and offers various other advantages over likelihood ratio tests [6].

## 2.3 Some other features of FINDMODEL

Finding the best evolutionary model is a computationally intensive procedure, both in its original implementation as the Modeltest PAUP script and in the FINDMODEL

Table 2.1: Models considered by FINDMODEL. Models in the reduced set are in bold and with reference.

Key	Model	# params	Ref.
<b>JC</b>	<b>Jukes-Cantor</b>	<b>0</b>	[16]
<b>F81</b>	<b>Felsenstein 81</b>	<b>3</b>	[8]
<b>K2P</b>	<b>Kimura 2-parameter</b>	<b>1</b>	[18]
<b>HKY</b>	<b>Hasegawa-Kishino-Yano</b>	<b>4</b>	[13]
TrNeq	Tamura-Nei equal-freq	2	
<b>TrN</b>	<b>Tamura-Nei</b>	<b>5</b>	[42]
K81	Kimura 3-parameter	2	
K81ne	Kimura 3p unequal-freq	5	
TIMEq	Transition equal-freq	3	
TIM	Transition	6	
TVMeq	Transversion equal-freq	4	
TVM	Transversion	7	
SYM	Symmetrical	5	
<b>GTR</b>	<b>General Time-reversible</b>	<b>8</b>	[34]

implementation.

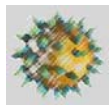
- FINDMODEL can find the best-fit model among twenty-eight models—see Table 2.1 (note that a  $\Gamma$  deviation can be added to every model). However, in order to reduce the computational burden on the server, a default run uses a reduced set of twelve models (models in bold font in Table 2.1 and those models with  $\Gamma$ ); the full set of models can be run as an option.
- Our running-time tests have been conducted on sequences of around 10,000 nucleotides. These tests show that, on the current server, FINDMODEL takes 24h to run for the full set of models on an input file of about 355kB or for the reduced set of models on an input file of about 520kB. Shorter sequences actually slow down the process as they require more iterations in the likelihood computations. Accordingly, we set a threshold of 350kB for the full set of models and 500kB for the reduced set of models for the maximum input size. (We are planning to release a down-loadable version that will enable users to run

on their own machines for as long as desired.)

- Input files larger than 3kB for the full set of models or 6kB for the reduced set of models may take over 5 minutes to complete. We used these sizes as thresholds to classify jobs as interactive or batch-mode. When the file size exceeds the threshold, the job is run in the background and the results stored for one week on the server from where they can be retrieved at an address provided in the email sent to the user upon completion of the analysis. A rough estimate of the anticipated running time is given before the job runs in this case.
- When the result is showed instantly on the web, FINDMODEL shows the parameter matrix for the selected model on its result page (Figure 2.2), and the parameter matrix for other models considered can be shown on the same page by clicking the model name.

FindModel input

<http://hcv.lanl.gov/content/hcv-db/findmodel/fin...>




**Find**


Housekeeping  
HCV classification  
[How to use this site](#)  
[Comments](#)  
[Help files](#)

Retrieving data  
[Search interface](#)  
[Geography](#)  
[Alignments](#)  
[Flaviviruses](#)  
[HCV variability](#)  
[External datasets](#)

Sequence analysis  
[TreeMaker](#)  
[HCV-BLAST](#)  
[Syn-mosaic](#)  
[PCOORD](#)  
[N-Glycosite](#)  
[Entropy](#)  
[Findmodel](#)

Tools  
[SynchAlkins](#)  
[Consensus](#)  
[Gene Cutter](#)  
[Sequence Locator](#)  
[OmniRead](#)  
[SeqConvert](#)  
[GapShrink/squeeze](#)  
[PeptGen](#)  
[MotifScan](#)  
[Primalpin](#)  
[EpiScan](#)  
[SeqPublish](#)

Links  
[Immunology DB](#)   
[Database software](#)  
[Conferences](#)  
[Web resources](#)  
[HIV databases](#)

 [link goes to immunology website](#)

[Disclaimer](#)

### HCV Sequence Database

#### FindModel

Purpose: Findmodel analyzes your alignment to see which phylogenetic model best describes your data; this model can then be used to generate a better tree.

Explanation: Findmodel was developed from a web implementation of the Modeltest script written by David Posada and Keith Crandall. It uses Bill Bruno's program [Weightbox](#) to generate the tree based on Jukes-Cantor distances. Weightbox is used because it is much faster than maximum likelihood, but less biased and more robust than NJ. Ziheng Yang's PAML is used to calculate the likelihood. The method from David Posada and Keith Crandall's [MODEL TEST paper](#) is used to calculate AIC scores. One difference to the Modeltest evaluation is that we do not allow invariant sites, as this feature is not implemented in PAML because estimates of the fraction of invariant sites tend to be very sensitive to the number or taxa.

Finding the best evolutionary model is a computationally intensive procedure, both in its original implementation as the Modeltest PAUP\* script and in our Findmodel implementation. To reduce the computational burden on our servers, we have limited the default runs to a reduced set of models, and excluded those that do not have an obvious biological interpretation. (If you know of any system where Modeltest consistently returns a model that we do not include, please let us know.) The full set of models can be run by checking the checkbox below the input section. Currently, input files smaller than 6 Kb for the reduced set and 3 Kb for the full set are run immediately; if your input file exceeds the limit, your job will be run in batch, and you will receive an email when it has finished. The email contains a link to your results. Currently, input files larger than 350 Kb for the reduced set and 500 Kb for the full set are too large for our machine to process.

How to use: Findmodel attempts to automatically recognize the format of your input file, using Format Conversion. If this fails, you can use the [Sequence Conversion Tools](#) interface or the [EMBOSS sequence conversion](#) site to convert your alignment to fasta, and use that for input.

[Sample Input](#)

[Browse...](#)

☐ use all the 28 models

[Submit](#)   [Reset](#)

This program is computationally intensive and may take a while to run; please don't resubmit your request!

Contributors to this implementation of findmodel include Ning Tao, Russell Richardson, William Bruno and Carla Kuiken.

For background information on selecting evolutionary models, see (for example): Posada D, Crandall KA. Selecting the best-fit model of nucleotide substitution. [Syst Biol 2001 Aug;50\(4\):580-601](#).

Johan Nylander has written an [implementation](#) of Modeltest that is based on [MrBayes](#) rather than PAUP.

Questions or comments? Contact us at [hcv-info@h10.lanl.gov](mailto:hcv-info@h10.lanl.gov)

This page was last updated on Apr 15, 2005.

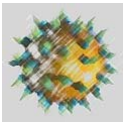


Operated by the University of California for the US Department of Energy  
 Copyright © 2001 LUC | [Disclaimer](#) | [Privacy](#)





Figure 2.1: The home page of FINDMODEL web site





[Find](#)

[Housekeeping](#)  
[HCV classification](#)  
[How to use this site](#)  
[Comments](#)  
[Help files](#)  
[Retrieving data](#)  
[Search interface](#)  
[Geography](#)  
[Alignments](#)  
[Flaviviruses](#)  
[HCV variability](#)  
[External datasets](#)  
[Sequence analysis](#)  
[TreeMaker](#)  
[HCV-BLAST](#)  
[Syn-nonsyn](#)  
[PCOORD](#)  
[N-Glycosite](#)  
[Entropy](#)  
[Findmodel](#)  
[Tools](#)  
[SynchAligns](#)  
[Consensus](#)  
[Gene Cutter](#)  
[Sequence Locator](#)  
[OmniRead](#)  
[Seq-Convert](#)  
[Gapstrip/squeeze](#)  
[PeptGen](#)  
[MotifScan](#)  
[Primalign](#)  
[Epsilon](#)  
[SeqPublish](#)  
[Links](#)  
[Immunology DB](#)   
[Database software](#)  
[Conferences](#)  
[Web resources](#)  
[HIV databases](#)  
 link goes to immunology website  
[Disclaimer](#)

## HCV Sequence Database

## FindModel Results

This program is computationally intensive and may take a while to run; please don't resubmit your request!

## MODEL CONSIDERED:

Model name	AIC	LnL
JC : Jukes-Cantor (model 1)	3563.252246	-1781.626123
JC+G : Jukes-Cantor plus Gamma (model 3)	3504.693372	-1751.346686
F81 : Felsenstein 1981 (model 5)	3564.440496	-1779.220248
F81+G : Felsenstein 1981 plus Gamma (model 7)	3502.851558	-1747.425779
K80 : Kimura 2-parameter (model 9)	3499.844658	-1748.922329
K80+G : Kimura 2-parameter plus Gamma (model 11)	3423.016278	-1709.508139
HKY : Hasegawa-Kishino-Yano (model 13)	3499.375768	-1745.687884
HKY+G : Hasegawa-Kishino-Yano plus Gamma (model 15)	3412.457576	-1701.228788
TrN : Tamura-Nei (model 21)	3494.642212	-1742.321106
TrN+G : Tamura-Nei plus Gamma (model 23)	3413.187698	-1700.593849
GTR : General Time Reversible (model 53)	3487.768892	-1735.884446
GTR+G : General Time Reversible plus Gamma (model 55)	3411.658894	-1696.829447

AIC-SELECTED MODEL: **GTR+G : General Time Reversible plus Gamma (model 55)**

LnL = -1696.829447

AIC = 3411.658894

For more information about the parameters, click [here](#).

These results are based on PAML likelihoods computed on a neighbor reference tree; they could be different from what would be obtained using PAUP.

Models with invariant sites are not supported in PAML and not considered here.

This result was generated at: Sat 4/23/2005 16:57

General Time Reversible +  $\gamma$ 

	T	C	A	G
T	$\pi_T$	a	b	c
C	a	$\pi_C$	d	e
A	b	d	$\pi_A$	f
G	c	e	f	$\pi_G$

(Rate Parameters with the same color are assumed to have the same value by the model.

Base Frequencies are all the same when they

Click one of the model names in the "Model Considered" table to show the parameter matrix for that model here

Questions or comments? Contact us at [hcv-info@t10.lanl.gov](mailto:hcv-info@t10.lanl.gov)

Figure 2.2: One of the result pages of FINDMODEL web site

# Chapter 3

## Experimental Setup

To test the performance of FINDMODEL, we generated sets of simulated DNA sequences under selected models of nucleotide substitution and compared the predictions made by FINDMODEL on these sets of sequences with the actual model used to generate them. Simulated data were generated with **Seq-Gen** 1.3.1 [33], which simulates the evolution of an “ancestral” sequence down the edges of a phylogenetic tree using any one of a large variety of models of nucleotide substitution. Relative state frequencies, transition to transversion ratio, and general reversible rate matrix may all be specified, as well as site-specific rate heterogeneity. The tree used was generated using Treemaker from the HCV database [21] with its sample input; this tree has 16 leaves and is relatively balanced. All tests were done with the reduced set of models.

We generated simulated DNA sequences under four different models chosen from the reduced set of models in Table 2.1: Jukes-Cantor (JC) [16], Kimura 2-parameter (K2P) [18], Hasegawa-Kishino-Yano (HKY) [13], and General Time-Reversible (GTR) [34] (see Table 3.1, Table 3.2, and Table 3.3). We chose these four models because they all have an obvious biological interpretation and because they span all complexity levels, from the simplest (JC) to the most complex (GTR). For each

Table 3.1: Parameters for tests of the JC and K2P models

model	test #	sequence length	transition to transversion
JC	1	329	
	2	1,000	
K2P	1	329	1.0
	2	329	2.0
	3	1,000	2.0

model, we simulated sequences of 329 and 1000 nucleotides in order to test the sensitivity of FINDMODEL to the length of the sequences. Transition-to-transversion ratios of 1.0 and 2.0 were used for models K2P and HKY. JC is a special case of K2P and corresponds to K2P with transition-to-transversion ratio of 0.5. The ratio 2.0 was used for more tests than 1.0, since it is closer to the real transition-to-transversion ratio for hepatitis C sequences—our original application. Relative state frequencies, which represents the equilibrium frequencies of the four nucleotides, were calculated from real hepatitis C or HIV sequences.

**Seq-Gen** implements site-specific rate heterogeneity, under which different sites evolve at different rates. A particularly simple way to specify such heterogeneity is to use a gamma distribution, usually considered the most appropriate approximation for rate differences among the variable sites [3]. A shape parameter,  $\alpha$ , for the  $\Gamma$  rate heterogeneity must be specified, with lower values denoting more variation across sites. Typical values estimated from real data tend to be around 0.3, so we used values of 0.1 and 0.5 in order to test the effect of large and small variations across sites.

Table 3.2: Parameters for tests of the HKY model

test #	sequence length	transition to transversion	A,C,G,T frequencies	discrete $\Gamma$	$\Gamma$
1	329	2.0	0.23366, 0.26786 0.29369, 0.20479	5	0.1
2	1,000	2.0	0.23366, 0.26786 0.29369, 0.20479		
3	1,000	2.0	0.19986, 0.29092 0.27624, 0.21329		
4	1,000	1.2	0.34357, 0.17696 0.23013, 0.24937		
5	1,000	2.0	0.19986, 0.29092 0.27624, 0.21329		
6	1,000	2.0	0.19986, 0.29092 0.27624, 0.21329		
7	1,000	2.0	0.19986, 0.29092 0.27624, 0.21329		0.5

Table 3.3: Parameters for tests of the GTR model

test #	sequence length	discrete $\Gamma$	$\Gamma$
1	329	5	0.1
2	1,000		
3	1,000		
4	1,000	5	0.5
5	1,000		0.5

	C	G	T
A	0.839597	0.083972	0.132634
C		0.177409	0.257970
G			0.579553

# Chapter 4

## Results

### 4.1 Results Regarding Models

Table 4.1 summarizes the results in terms of model matches and mismatches; we report the results separately for each of the 5 instances generated for each of the 17 distinct groups of parameters.

These results can be viewed as follows:

- JC, the most specific model, was selected in 6 out of 10 test instances for JC (see Table 4.2 and Table 4.3); of the other four test instances, three selected a more general model, one selected JC+ $\Gamma$ .
- K2P was selected in 12 out of 15 test instances for K2P (see Table 4.4 to Table 4.6); of the other three test instances, only one selected a more general model (HKY+ $\Gamma$ ), while two selected K2P+ $\Gamma$ .
- HKY was selected in 11 out of 20 test instances for HKY (see Table 4.7 to Table 4.10); of the other nine test instances, one selected a more general model (GTR), one selected a more specific model (K2P), four selected TrN (very

Table 4.1: Test results. Empty entries indicate perfect matches; an asterisk (\*) indicates a match, but with a  $\Gamma$  rate parameter added; a question mark (?) denotes a mismatch of low significance, where the AIC error rate (*AIC difference between correct model and selected model / AIC of selected model*) was below 0.0005 (all entries marked with an asterisk also met this criterion); finally, more significant mismatches are indicated by the erroneous choice of model and, in parentheses, the corresponding AIC error rate. (The AIC values obtained for each test are shown in Table 4.2 to Table 4.18)

Model tested	test #	1	2	3	4	5
JC	1	K2P (0.001127)			*	?
	2					?
K2P	1					
	2			?	*	
	3	*				
HKY	1	?	?		TrN (0.000863)	
	2		?			*
	3			?		
HKY+ $\Gamma$	4		?	*		*
	5		?	?		
	6				?	
	7	?				?
GTR	1	*				
	2				*	
GTR+ $\Gamma$	3					
	4					
	5					

similar to HKY [44]), while three selected HKY+ $\Gamma$ .

- HKY+ $\Gamma$  was selected in 10 out of 15 tests for HKY+ $\Gamma$  (see Table 4.11 to Table 4.13); of the other five test instances, three selected a more general model (GTR+ $\Gamma$ ) and two selected TrN+ $\Gamma$  (very similar to HKY+ $\Gamma$  [44]).
- GTR was selected in 8 out of 10 test instances for GTR (see Table 4.14 and Table 4.15), while the other two test instances selected GTR+ $\Gamma$ .
- GTR+ $\Gamma$  was selected in all 15 test instances for GTR+ $\Gamma$  (see Table 4.16 to Table 4.18).

Table 4.2: AIC values of JC model test 1. In this and all the following AIC value tables, the AIC values of selected models are in bold, and the ones of correct models are in italic.

model	run 1	run 2	run 3	run 4	run 5
JC(1)	<i>6187.173</i>	<b>5613.264</b>	<b>6039.056</b>	<i>5918.637</i>	<i>6071.035</i>
JC+G(3)	6189.148	5614.764	6040.734	<b>5916.842</b>	6072.443
F81(5)	6191.699	5617.861	6044.925	5925.017	6075.903
F81+G(7)	6193.646	5619.380	6046.506	5923.173	6077.342
K80(9)	<b>6180.210</b>	5615.193	6040.797	5920.491	<b>6070.789</b>
K80+G(11)	6182.183	5616.696	6042.466	5918.685	6072.196
HKY(13)	6184.625	5619.803	6046.708	5926.857	6075.712
HKY+G(15)	6186.568	5621.325	6048.282	5925.000	6077.144
TrN(21)	6186.620	5621.799	6047.150	5928.773	6076.631
TrN+G(23)	6188.561	5623.321	6048.746	5926.896	6078.131
GTR(53)	6191.983	5625.075	6049.560	5934.354	6082.488
GTR+G(55)	6193.943	5626.648	6051.186	5932.443	6083.962
Model selected	K80	JC	JC	JC+G	K80

Overall, among 85 test instances, 62 test instances selected the correct models and an additional 8 selected the correct model plus  $\Gamma$ ; in the latter case, the AIC score

Table 4.3: AIC values of JC model test 2

model	run 1	run 2	run 3	run 4	run 5
JC(1)	<b>17904.041</b>	<b>17928.909</b>	<b>17953.130</b>	<b>18175.153</b>	<i>18058.687</i>
JC+G(3)	17905.993	17929.547	17955.133	18176.841	18060.733
F81(5)	17910.383	17931.926	17955.698	18179.151	<b>18058.026</b>
F81+G(7)	17912.329	17932.571	17957.699	18180.842	18060.076
K80(9)	17905.692	17930.076	17954.980	18175.214	18060.528
K80+G(11)	17907.643	17930.713	17956.983	18176.909	18062.574
HKY(13)	17912.038	17933.092	17957.507	18179.172	18059.832
HKY+G(15)	17913.983	17933.733	17959.508	18180.871	18061.881
TrN(21)	17909.502	17934.734	17959.494	18180.914	18061.394
TrN+G(23)	17911.463	17935.391	17961.495	18182.617	18063.446
GTR(53)	17914.985	17938.911	17957.487	18185.412	18063.569
GTR+G(55)	17916.949	17939.565	17959.490	18187.140	18065.620
Model selected	JC	JC	JC	JC	F81

Table 4.4: AIC values of K2P model test 1

model	run 1	run 2	run 3	run 4	run 5
JC(1)	5946.753	5897.106	5731.961	5762.515	5982.065
JC+G(3)	5948.255	5899.168	5733.999	5764.532	5984.090
F81(5)	5953.307	5897.493	5736.867	5765.723	5988.037
F81+G(7)	5954.786	5899.555	5738.907	5767.739	5990.062
K80(9)	<b>5893.216</b>	<b>5815.355</b>	<b>5659.919</b>	<b>5691.343</b>	<b>5954.134</b>
K80+G(11)	5894.539	5817.409	5661.951	5693.354	5956.157
HKY(13)	5899.527	5818.641	5663.459	5692.457	5959.708
HKY+G(15)	5900.816	5820.696	5665.493	5694.467	5961.730
TrN(21)	5900.843	5820.220	5663.885	5692.456	5961.243
TrN+G(23)	5902.120	5822.275	5665.920	5694.468	5963.266
GTR(53)	5905.032	5819.519	5665.277	5693.984	5958.963
GTR+G(55)	5906.186	5821.572	5667.312	5695.995	5960.988
Model selected	K80	K80	K80	K80	K80

of the chosen model was within 0.05% of that the correct model (a difference of less than 3). Among the remaining 15 test instances, the second best model was the correct one in 10 test instances and always with an AIC score within 0.12% of

Table 4.5: AIC values of K2P model test 2

model	run 1	run 2	run 3	run 4	run 5
JC(1)	5978.390	5757.453	5882.386	5966.160	5722.158
JC+G(3)	5980.034	5757.668	5884.430	5966.510	5724.169
F81(5)	5983.770	5756.708	5888.410	5971.494	5728.835
F81+G(7)	5985.431	5757.028	5890.453	5971.902	5730.847
K80(9)	<b>5741.613</b>	<i>5515.339</i>	<b>5658.399</b>	<i>5720.481</i>	<b>5487.847</b>
K80+G(11)	5741.895	5514.191	5660.431	<b>5719.430</b>	5489.832
HKY(13)	5745.629	5514.161	5664.474	5726.679	5494.772
HKY+G(15)	5745.914	<b>5513.396</b>	5666.506	5725.970	5496.744
TrN(21)	5747.629	5516.128	5666.474	5728.355	5493.427
TrN+G(23)	5747.914	5515.378	5668.506	5727.567	5495.419
GTR(53)	5745.555	5518.917	5671.610	5725.726	5497.689
GTR+G(55)	5745.490	5518.306	5673.642	5725.266	5499.686
Model selected	K80	HKY+G	K80	K80+G	K80



Table 4.6: AIC values of K2P model test 3

model	run 1	run 2	run 3	run 4	run 5
JC(1)	17816.704	17603.196	17796.847	18005.106	17972.443
JC+G(3)	17818.041	17605.268	17798.909	18007.208	17974.503
F81(5)	17820.103	17610.036	17794.014	18007.876	17977.693
F81+G(7)	17821.410	17612.108	17796.077	18009.976	17979.755
K80(9)	<i>17114.467</i>	<b>16909.621</b>	<b>16983.106</b>	<b>17311.731</b>	<b>17236.817</b>
K80+G(11)	<b>17113.753</b>	16911.652	16985.129	17313.790	17238.833
HKY(13)	17117.660	16915.969	16983.673	17315.696	17241.620
HKY+G(15)	17116.814	16917.998	16985.697	17317.756	17243.637
TrN(21)	17119.628	16917.961	16985.041	17317.686	17242.945
TrN+G(23)	17118.790	16919.991	16987.065	17319.745	17244.964
GTR(53)	17124.080	16921.402	16987.403	17320.071	17244.951
GTR+G(55)	17123.361	16923.433	16989.424	17322.133	17246.970
Model selected	K80+G	K80	K80	K80	K80

that of the correct model (a difference of less than 7). In only one test instance did FINDMODEL choose a model more specific than the correct model, while it chose a more general model in 14 test instances.

Table 4.7: AIC values of HKY model test 1

model	run 1	run 2	run 3	run 4	run 5
JC(1)	5957.457	5915.157	5795.688	5688.362	6250.941
JC+G(3)	5959.483	5917.007	5797.691	5689.487	6252.537
F81(5)	5961.965	5913.734	5775.481	5677.574	6241.887
F81+G(7)	5963.991	5915.641	5777.481	5678.672	6243.725
K80(9)	<b>5741.231</b>	5728.215	5565.495	5504.463	6004.744
K80+G(11)	5743.249	5729.683	5567.370	5504.820	6005.742
HKY(13)	<i>5741.569</i>	<i>5727.019</i>	<b>5548.728</b>	<i>5500.918</i>	<b>5994.275</b>
HKY+G(15)	5743.588	5728.481	5550.727	5500.637	5995.877
TrN(21)	5742.425	<b>5725.486</b>	5550.572	<b>5496.175</b>	5996.258
TrN+G(23)	5744.446	5727.039	5552.573	5496.239	5997.854
GTR(53)	5745.370	5726.442	5556.319	5500.252	6000.811
GTR+G(55)	5747.391	5727.969	5558.319	5500.218	6002.488
Model selected	K80	TrN	HKY	TrN	HKY

Table 4.8: AIC values of HKY model test 2

model	run 1	run 2	run 3	run 4	run 5
JC(1)	17121.935	17830.410	18151.487	18212.984	17790.584
JC+G(3)	17123.900	17832.463	18153.616	18214.986	17790.754
F81(5)	17085.530	17775.801	18111.336	18191.429	17761.697
F81+G(7)	17087.473	17777.864	18113.464	18193.439	17762.461
K80(9)	16549.344	17120.804	17383.076	17464.740	17037.068
K80+G(11)	16550.634	17122.807	17385.159	17466.027	17033.336
HKY(13)	<b>16497.343</b>	<i>17066.590</i>	<b>17347.505</b>	<b>17421.998</b>	<i>16997.629</i>
HKY+G(15)	16498.527	17068.606	17349.587	17423.570	<b>16995.012</b>
TrN(21)	16498.761	17067.271	17348.741	17423.998	16997.365
TrN+G(23)	16499.960	17069.289	17350.825	17425.570	16995.062
GTR(53)	16503.271	<b>17064.358</b>	17353.379	17429.522	17002.269
GTR+G(55)	16504.491	17066.375	17355.462	17431.074	16999.979
Model selected	HKY	GTR	HKY	HKY	HKY+G

## 4.2 Results Regarding Lengths of Sequences

- For the JC model, JC was selected in 4 out of 5 test instances for sequences of length of 1000 (see Table 4.3), but in 2 out of 5 test instances for sequences of

Table 4.9: AIC values of HKY model test 3

model	run 1	run 2	run 3	run 4	run 5
JC(1)	17355.857	17103.030	17241.481	17510.880	17347.000
JC+G(3)	17357.953	17105.114	17243.488	17510.878	17349.111
F81(5)	17324.741	17058.837	17234.789	17479.883	17314.017
F81+G(7)	17326.843	17060.947	17236.808	17481.019	17316.148
K80(9)	16730.545	16367.574	16588.695	16860.437	16753.217
K80+G(11)	16732.600	16369.612	16590.277	16858.198	16755.291
HKY(13)	<b>16675.588</b>	<b>16275.207</b>	<i>16555.544</i>	<b>16799.870</b>	<b>16698.583</b>
HKY+G(15)	16677.645	16277.278	16557.284	16800.007	16700.683
TrN(21)	16677.558	16277.117	<b>16554.535</b>	16801.406	16700.581
TrN+G(23)	16679.615	16279.188	16556.318	16801.532	16702.681
GTR(53)	16681.913	16281.135	16558.097	16799.982	16705.895
GTR+G(55)	16683.971	16283.207	16559.872	16800.218	16707.996
Model selected	HKY	HKY	TrN	HKY	HKY

Table 4.10: AIC values of HKY model test 4

model	run 1	run 2	run 3	run 4	run 5
JC(1)	17366.278	17961.585	18149.752	17335.909	17541.899
JC+G(3)	17368.135	17962.289	18149.061	17337.968	17540.799
F81(5)	17245.188	17841.448	17984.275	17150.599	17385.382
F81+G(7)	17247.146	17842.617	17984.410	17152.653	17383.373
K80(9)	17089.823	17719.663	17852.865	17060.594	17284.593
K80+G(11)	17091.306	17719.587	17850.986	17062.639	17282.422
HKY(13)	<b>16949.074</b>	<i>17585.866</i>	<i>17699.647</i>	<b>16893.054</b>	<i>17155.608</i>
HKY+G(15)	16950.800	17586.817	<b>17699.443</b>	16895.093	<b>17154.139</b>
TrN(21)	16949.220	<b>17585.467</b>	17701.065	16894.300	17156.963
TrN+G(23)	16950.878	17586.408	17700.892	16896.340	17155.353
GTR(53)	16951.331	17590.508	17706.254	16895.771	17160.032
GTR+G(55)	16952.993	17591.466	17706.108	16897.814	17158.520
Model selected	HKY	TrN	HKY+G	HKY	HKY+G

length of 329 (see Table 4.2). For the test instances that did not select JC, the AIC values of JC were closer to the AIC values of selected model for sequences of length 1000 (see values in Table 4.1 and Table 3.1).

Table 4.11: AIC values of HKY+ $\Gamma$  model test 5

model	run 1	run 2	run 3	run 4	run 5
JC(1)	16609.895	16638.548	16156.544	16726.336	15672.367
JC+G(3)	16331.209	16292.661	15926.510	16515.682	15423.283
F81(5)	16575.938	16610.542	16138.605	16688.679	15633.872
F81+G(7)	16297.049	16260.539	15914.561	16483.798	15379.330
K80(9)	16149.007	16130.809	15611.691	16194.630	15199.554
K80+G(11)	15846.640	15749.465	15358.825	15958.079	14924.765
HKY(13)	16086.423	16094.524	15565.960	16135.910	15130.965
HKY+G(15)	<b>15781.990</b>	<i>15707.910</i>	<i>15322.691</i>	<b>15907.849</b>	<b>14849.688</b>
TrN(21)	16087.985	16094.851	15564.879	16136.585	15132.614
TrN+G(23)	15783.576	<b>15707.709</b>	15323.168	15908.614	14850.297
GTR(53)	16093.942	16100.119	15557.626	16139.237	15135.006
GTR+G(55)	15789.414	15713.446	<b>15318.741</b>	15910.379	14852.300
Model selected	HKY+G	TrN+G	GTR+G	HKY+G	HKY+G

Table 4.12: AIC values of HKY+ $\Gamma$  model test 6

model	run 1	run 2	run 3	run 4	run 5
JC(1)	10998.935	11531.007	11121.435	11527.585	11625.601
JC+G(3)	9646.508	10049.084	9909.204	10028.613	10102.357
F81(5)	10959.929	11502.727	11092.874	11499.275	11589.851
F81+G(7)	9607.761	10013.216	9883.974	9986.241	10065.136
K80(9)	10898.239	11356.353	10948.290	11412.032	11465.837
K80+G(11)	9517.631	9831.340	9704.927	9891.021	9909.537
HKY(13)	10849.764	11317.880	10905.683	11370.840	11417.543
HKY+G(15)	<b>9470.315</b>	<b>9780.490</b>	<b>9666.172</b>	<i>9828.112</i>	<b>9861.773</b>
TrN(21)	10851.396	11314.633	10907.665	11372.787	11418.381
TrN+G(23)	9470.750	9781.626	9668.168	9829.096	9863.769
GTR(53)	10854.643	11314.551	10906.821	11377.013	11418.092
GTR+G(55)	9474.674	9783.169	9668.063	<b>9828.020</b>	9868.693
Model selected	HKY+G	HKY+G	HKY+G	GTR+G	HKY+G

- For the K2P model, when the transition-to-transversion ration was 2.0, K2P was selected in 4 out of 5 test instances for sequences of length 1000 (see Table 4.6), but in 3 out of 5 test instances for sequences of length of 329 (see

Table 4.13: AIC values of HKY+ $\Gamma$  model test 7

model	run 1	run 2	run 3	run 4	run 5
JC(1)	15817.025	15206.093	15514.700	15259.629	15248.885
JC+G(3)	15108.967	14614.544	14856.303	14552.872	14650.124
F81(5)	15796.269	15177.213	15468.932	15237.713	15220.111
F81+G(7)	15086.428	14584.418	14820.619	14529.958	14630.665
K80(9)	15481.583	14762.347	15141.435	14944.625	14894.555
K80+G(11)	14732.193	14128.019	14439.531	14206.078	14267.123
HKY(13)	15442.492	14718.672	15060.023	14909.092	14838.147
HKY+G(15)	<i>14686.217</i>	<b>14086.825</b>	<b>14374.292</b>	<b>14167.152</b>	<i>14221.549</i>
TrN(21)	15442.074	14720.240	15061.784	14909.864	14840.092
TrN+G(23)	<b>14685.134</b>	14087.547	14376.288	14169.126	14223.267
GTR(53)	15446.048	14724.735	15061.200	14912.855	14829.838
GTR+G(55)	14690.528	14091.385	14377.594	14170.915	<b>14215.511</b>
Model selected	TrN+G	HKY+G	HKY+G	HKY+G	GTR+G

Table 4.14: AIC values of GTR model test 1

model	run 1	run 2	run 3	run 4	run 5
JC(1)	5849.225	6082.968	5891.277	5812.723	6009.343
JC+G(3)	5847.193	6084.480	5893.273	5814.716	6011.341
F81(5)	5845.736	6079.250	5902.463	5813.331	6019.380
F81+G(7)	5843.089	6080.444	5904.465	5815.303	6021.381
K80(9)	5782.763	5999.218	5835.064	5736.894	5917.601
K80+G(11)	5780.147	6000.551	5837.030	5738.873	5919.567
HKY(13)	5779.635	5994.507	5846.416	5737.707	5926.550
HKY+G(15)	5776.525	5995.467	5848.414	5739.653	5928.545
TrN(21)	5748.425	5984.100	5802.035	5726.205	5910.186
TrN+G(23)	5747.523	5985.461	5804.057	5728.186	5912.187
GTR(53)	<i>5562.086</i>	<b>5743.932</b>	<b>5597.629</b>	<b>5566.414</b>	<b>5741.158</b>
GTR+G(55)	<b>5561.026</b>	5745.355	5599.654	5568.408	5742.976
Model selected	GTR+G	GTR	GTR	GTR	GTR

Table 4.5). For the test instances that did not select K2P, the AIC values of K2P were closer to the AIC values of selected model for sequences of length 1000 (see values in Table 4.1 and Table 3.1).

Table 4.15: AIC values of GTR model test 2

model	run 1	run 2	run 3	run 4	run 5
JC(1)	17682.316	17314.484	17838.932	17626.886	17451.606
JC+G(3)	17684.404	17316.526	17841.010	17624.766	17453.643
F81(5)	17681.538	17309.469	17825.300	17635.881	17465.204
F81+G(7)	17683.622	17311.508	17827.371	17633.907	17467.246
K80(9)	17474.072	17115.157	17656.029	17426.513	17217.537
K80+G(11)	17476.140	17117.190	17658.097	17423.166	17219.545
HKY(13)	17476.087	17110.295	17642.593	17435.653	17229.835
HKY+G(15)	17478.148	17112.325	17644.653	17432.468	17231.851
TrN(21)	17432.737	17059.981	17582.352	17391.896	17174.644
TrN+G(23)	17434.812	17062.010	17584.436	17389.632	17176.681
GTR(53)	<b>16817.843</b>	<b>16520.270</b>	<b>17034.174</b>	<i>16907.388</i>	<b>16641.655</b>
GTR+G(55)	16819.908	16522.285	17036.236	<b>16905.011</b>	16643.672
Model selected	GTR	GTR	GTR	GTR+G	GTR

Table 4.16: AIC values of GTR+ $\Gamma$  model test 3

model	run 1	run 2	run 3	run 4	run 5
JC(1)	16682.116	16514.702	16702.632	16668.498	17040.233
JC+G(3)	16371.438	16193.319	16371.652	16346.523	16696.926
F81(5)	16690.981	16510.882	16707.304	16671.928	17058.894
F81+G(7)	16380.509	16187.451	16377.870	16348.538	16717.948
K80(9)	16495.662	16372.640	16548.909	16503.640	16844.615
K80+G(11)	16178.137	16044.080	16207.471	16173.254	16490.586
HKY(13)	16503.991	16368.937	16553.558	16507.322	16860.808
HKY+G(15)	16186.684	16039.050	16213.727	16175.899	16510.534
TrN(21)	16445.224	16336.897	16506.088	16454.258	16794.558
TrN+G(23)	16121.378	16004.067	16174.621	16129.803	16450.634
GTR(53)	16019.615	15914.206	16061.610	15960.335	16331.511
GTR+G(55)	<b>15681.901</b>	<b>15560.559</b>	<b>15700.179</b>	<b>15620.379</b>	<b>15960.109</b>
Model selected	GTR+G	GTR+G	GTR+G	GTR+G	GTR+G

- For the HKY model, when the transition-to-transversion ratio was 2.0 and the values for relative state frequencies were the same (0.23366, 0.26786, 0.29369, and 0.20479), HKY was selected in 3 out of 5 test instances for sequences of

Table 4.17: AIC values of GTR+ $\Gamma$  model test 4

model	run 1	run 2	run 3	run 4	run 5
JC(1)	11441.426	12592.245	12517.206	12204.339	11716.968
JC+G(3)	9951.129	10913.741	10867.033	10494.368	9974.928
F81(5)	11442.756	12597.116	12510.782	12201.795	11722.043
F81+G(7)	9950.976	10920.671	10861.069	10492.379	9989.805
K80(9)	11385.025	12521.564	12466.575	12142.684	11674.845
K80+G(11)	9884.831	10829.993	10804.839	10403.529	9911.281
HKY(13)	11386.452	12527.104	12460.108	12140.466	11675.037
HKY+G(15)	9885.689	10836.718	10798.811	10402.419	9927.859
TrN(21)	11370.055	12508.469	12452.269	12142.229	11621.400
TrN+G(23)	9875.806	10808.633	10782.134	10402.244	9892.902
GTR(53)	11214.222	12373.305	12271.197	12002.312	11519.300
GTR+G(55)	<b>9694.806</b>	<b>10622.844</b>	<b>10559.997</b>	<b>10212.349</b>	<b>9761.921</b>
Model selected	GTR+G	GTR+G	GTR+G	GTR+G	GTR+G

Table 4.18: AIC values of GTR+ $\Gamma$  model test 5

model	run 1	run 2	run 3	run 4	run 5
JC(1)	15457.691	15497.331	16017.649	15309.959	15792.612
JC+G(3)	14783.766	14915.423	15368.558	14629.768	15206.930
F81(5)	15459.070	15502.334	16031.353	15316.512	15797.457
F81+G(7)	14783.046	14919.942	15383.299	14636.858	15212.745
K80(9)	15345.557	15383.954	15866.676	15176.852	15677.371
K80+G(11)	14661.326	14788.138	15212.892	14486.218	15081.618
HKY(13)	15347.237	15388.997	15880.260	15183.449	15683.216
HKY+G(15)	14662.118	14792.908	15227.713	14493.223	15087.649
TrN(21)	15301.557	15349.933	15843.497	15149.211	15665.182
TrN+G(23)	14619.207	14747.261	15194.169	14458.660	15066.402
GTR(53)	14949.957	14973.356	15493.305	14773.869	15326.244
GTR+G(55)	<b><i>14231.609</i></b>	<b><i>14359.753</i></b>	<b><i>14809.274</i></b>	<b><i>14047.406</i></b>	<b><i>14701.597</i></b>
Model selected	GTR+G	GTR+G	GTR+G	GTR+G	GTR+G

length 1000 (see Table 4.8), but in 2 out of 5 test instances for sequences of length 329 (see Table 4.7, Table 4.1 and Table 3.2).

- Model GTR was correctly identified in 4 out of 5 test instances for sequences of either length (see Table 4.14 and Table 4.15). For the one test instance that did not select GTR, the AIC value of GTR was closer to the AIC value of the selected model for sequences of length 1000 (see values in Table 4.1 and Table 3.3).

These tests show that FINDMODEL results are more accurate when the sequences are longer, a common finding in phylogenetic analysis [3, 27].

### 4.3 Results Regarding $\Gamma$ Rate Heterogeneity

We used  $\Gamma$  rate heterogeneity in the testing for two models, HKY and GTR.

- For HKY, AIC values for models with  $\Gamma$  were much smaller than AIC values for the corresponding homogeneous model versions; they differed by about 2%

with 5 categories for the discrete  $\Gamma$  rate heterogeneity (see Table 4.11), by about 15% for  $\Gamma = 0.1$  (see Table 4.12), and by about 5% for  $\Gamma = 0.5$  (see Table 4.13). AIC values were smaller than AIC values in test instances using the same set of parameters but without  $\Gamma$  (see Table 4.9); they differed by about 6% for HKY test5, by about 36% for HKY test6, and by about 12% for HKY test7.

- For GTR, AIC values for models with  $\Gamma$  were much smaller than AIC values for the corresponding homogeneous model versions; they differed by about 2% with 5 categories for the discrete  $\Gamma$  rate heterogeneity (see Table 4.16), by 16% for  $\Gamma = 0.1$  (see Table 4.17), and by 5% for  $\Gamma = 0.5$  (see Table 4.18). AIC values were smaller than those in test instances using the same parameters without the  $\Gamma$  rate heterogeneity (see Table 4.15); they differed by about 6% for GTR test3, by about 40% for GTR test4, and by about 9% for GTR test5.

The difference between AIC value for models with and without  $\Gamma$  is much larger for  $\Gamma = 0.1$  than for  $\Gamma = 0.5$ , as one would expect (recall that a smaller  $\Gamma$  means more variation in rates). Thus FINDMODEL works correctly with  $\Gamma$ .



# Chapter 5

## Discussion

Many tools exist that can generate trees from sequence alignments, carry out phylogenetic analysis, generate likelihood values for edge parameters, and finally select the best-fit model of nucleotide substitution. However, most of these steps benefit from expert human intervention, something that nonspecialists in phylogenetic reconstruction may find intimidating. FINDMODEL runs on all platforms, provides a user-friendly interface, and carries out all steps of the analysis automatically, with well matched and statistically sound methods.

Our tests also show that FINDMODEL results are quite accurate, since it chose the correct model (sometimes plus  $\Gamma$ ) in 82% of the cases and since the AIC error rates in the remaining cases are always below 0.12% (and mostly below 0.05%). Among the 15 test instances in which the correct model was not chosen, 1 test instance chose a model more specific than the correct model, while 14 test instances chose a more general model.

FINDMODEL uses Weighbor to reconstruct a tree to feed to PAML. Other applications could be used, although both ML and Bayesian methods are currently unable to handle datasets of substantial size, while MP methods (which have been scaled to

20,000 sequences [35]) typically return large numbers of equally scoring trees, whose use in model finding remains to be determined. Currently FINDMODEL is targeted at datasets of less than 1,000 sequences, a range in which Weighbor usually works well.

Finding the best evolutionary model is a computationally intensive procedure. FINDMODEL is trivially parallelizable to a modest degree, since each separate model can be evaluated independently. More significantly, a web server will always remain bounded by the capacity of its hardware and the size of its customer base, so we are planning to release a down-loadable code that will allow users to run as long as desired on their own machines, using whichever phylogenetic packages they prefer.

# References

- [1] J. Adachi and M. Hasegawa. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. *J. Mol. Evol.*, 40(6):622–628, 1995.
- [2] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19:716–723, 1974.
- [3] D.H. Bos and D. Posada. Using models of nucleotide evolution to build phylogenetic trees. *Dev. Comp. Immunol.*, 29:211–227, 2005.
- [4] W.J. Bruno and A.L. Halpern. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.*, 16:564–566, 1999.
- [5] W.J. Bruno, N.D. Succi, and A.L. Halpern. Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, 17(1):189–197, 2000.
- [6] K.P. Burnham and D.R. Anderson. *Model selection and multimodel inference: A practical information-theoretic approach*. Springer Verlag, 2002.
- [7] C.W. Cunningham, H. Zhu, and D.M. Hillis. Best-fit maximum likelihood models for phylogenetic inference: Empirical tests with known phylogenies. *Evol.*, 52:978–987, 1998.
- [8] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.
- [9] J. Felsenstein. PHYLIP—phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.
- [10] J. Felsenstein. *Inferring Phylogenies*. Sinauer Assoc., Sunderland, MA, 2003.

## References

- [11] D. Graur and W.-H. Li. *Fundamentals of Molecular Evolution*. Sinauer Assoc., Sunderland, MA, 1991.
- [12] M. Hasegawa. Phylogeny and molecular evolution in primates. *Jap. J. Genetics*, 65(4):243–266, 1990.
- [13] M. Hasegawa, H. Kishino, and T.-A. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 21:160–174, 1985.
- [14] D.M. Hillis, M.W. Allard, and M.M. Miyamoto. Analysis of DNA sequence data: Phylogenetic inference. *Meth. Enzymol.*, 242:456–487, 1993.
- [15] D.M. Hillis, J.P. Huelsenbeck, and C.W. Cunningham. Application and accuracy of molecular phylogenies. *Science*, 264:671–677, 1994.
- [16] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.
- [17] C.R. Kelsey, K.A. Crandall, and A.F. Voevodin. Different models, different trees: the geographic origin of PTLV-I. *Mol. Phys. Evol.*, 13(2):336–347, 1999.
- [18] M. Kimura. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120, 1980.
- [19] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data. *J. Mol. Evol.*, 29:170–179, 1989.
- [20] S. Kosakovsky-Pond and S.V. Muse. *Hy-Phy: A Platform for Multilocus Molecular Evolutionary Analyses v0.9*, 2002. <http://peppercat.statgen.ncsu.edu/~hyphy>.
- [21] C. Kuiken, K. Yusim, L. Boykin, and R. Richardson. The Los Alamos hepatitis C sequence database. *Bioinformatics*, 21(3):379–384, 2005.
- [22] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, 1984.
- [23] T. Leitner, S. Kumar, and J. Albert. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *Journal of Virology*, 71:4761–4770, 1997.

- [24] W.-H. Li. *Molecular Evolution*. Sinauer Assoc., Sunderland, MA, 1997.
- [25] P. Lio and N. Goldman. Models of molecular evolution and phylogeny. *Genome Research*, 8:1233–1244, 1998.
- [26] B.M.E. Moret. Computational challenges from the Tree of Life. In *Proc. 7th SIAM Workshop on Algorithm Engineering & Experiments (ALENEX'05)*. SIAM Press, Philadelphia, 2005.
- [27] B.M.E. Moret, U. Roshan, and T. Warnow. Sequence length requirements for phylogenetic methods. In *Proc. 2nd Int'l Workshop Algs. in Bioinformatics (WABI'02)*, volume 2452 of *Lecture Notes in Computer Science*, pages 343–356. Springer Verlag, 2002.
- [28] M. Nei. Phylogenetic analysis in molecular evolutionary genetics. *Ann. Rev. Genetics*, 30:371–403, 1996.
- [29] R.D.M. Page and E.C. Holmes. *Molecular Evolution - A Phylogenetic Approach*. Blackwell Science, 1998.
- [30] D. Posada and K.A. Crandall. ModelTest: Testing the model of DNA substitution. *Bioinformatics*, 14(9):817–818, 1998.
- [31] D. Posada and K.A. Crandall. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc. Nat'l Acad. Sci., USA*, 98:13757–13762, 2001.
- [32] D. Posada and K.A. Crandall. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.*, 50(4):580–601, 2001.
- [33] A. Rambaut and N.C. Grassly. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13(3):235–238, 1997.
- [34] F. Rodriguez, J.L. Oliver, A. Marin, and J.R. Medina. The general stochastic model of nucleotide substitution. *J. Theor. Biol.*, 142(4):485–501, 1990.
- [35] U. Roshan, B.M.E. Moret, T.L. Williams, and T. Warnow. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. In *Proc. 3rd IEEE Computational Systems Bioinformatics Conf. CSB'04*, pages 98–109. IEEE Press, Piscataway, NJ, 2004.
- [36] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

- [37] J. Sullivan and D.L. Swofford. Are guinea pigs rodents? The importance of adequate models in molecular phylogenies. *J. Mammal. Evol.*, 4:77–86, 1997.
- [38] J. Sullivan and D.L. Swofford. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution process are violated? *Syst. Biol.*, 50:723–729, 2001.
- [39] D.L. Swofford. *PAUP\*: Phylogenetic analysis using parsimony (\*and other methods)*, version 4.0b8, 2001.
- [40] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, B.K. Mable, and C. Moritz, editors, *Molecular Systematics*, pages 407–514. Sinauer Assoc., Sunderland, MA, 1996.
- [41] K. Tamura. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. *Mol. Biol. Evol.*, 9:678–687, 1992.
- [42] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10(3):512–526, 1993.
- [43] J. Wakeley. Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.*, 11(3):436–442, 1994.
- [44] Z. Yang. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, 39(1):105–111, 1994.
- [45] Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39(3):306–314, 1994.
- [46] Z. Yang, N. Goldman, and A. Friday. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Syst. Biol.*, 44:384–399, 1995.
- [47] J. Zhang. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol. Biol. Evol.*, 16(6):868–875, 1999.